2

ARTIFICIAL INTELLIGENCE LABORATORY
and
LABORATORY FOR COMPUTER SCIENCE
MASSACHUSETTS INSTITUTE OF TECHNOLOGY

" A Computational Analysis of Properties and Limitations of Neural Networks: Toward New Parallel Architectures for Learning"

by T.A. Poggio and R. Rivest

Final Report

Technical contact: Prof. Tomaso Poggio, (617)253-5230, tp@ai.mit.edu

Administrative contact: Eileen Nielsen, (617)253-3491, eileen@ai.mit.edu

M.I.T. Artificial Intelligence Laboratory
545 Technology Square
Cambridge, MA 02139

Contents

1 Summary

The goal of our work has been to develop a solid theoretical framework for the problem of learning from examples, in order to evaluate Neural Network architectures and develop new powerful parallel techniques and algorithms. Our approach was based on the formulation of the problem of learning from examples as a problem of approximation of multivariate functions from sparse data, in such a way as to take advantage of existing large body of results in function approximation theory and regularization. Our work has been successfull beyond our original expectations at the time we wrote the proposal. We have developed a sizable body of theoretical results and applications. Several projects, many outside our own group, are now pursuing different aspects of the theory, and are develcping algorithms and applying the technique to practical domains. Below are some of the specific accomplishments:

- Theory of regularization networks
- Regularization networks contain RBF as a special case
- Extension of the theory: moving centers and task-dependent clustering
- Extension of the theory: moving centers and task-dependent dimensionality reduction
- A new optimization algorithm (for learning) and its parallel implementation on the Connection Machine
- Theory and numerical experiments on the relation HBF-multilayer perceptrons
- Sample complexity and function spaces
- Demonstration of new techniques for the following applications:
  - 3-D object recognition
  - Hyperacuity
  - Autonomous indoor navigation
  - Real 3-D object recognition
  - Time-series forecasting.

2 Where We Are Today

2.1 Present Directions of Research

The approach to learning that we have been developing over the past two years regards learning from examples as a problem of approximating a multivariate function from sparse data – the examples. We have developed a technique that has its roots in the classical theory of function approximation, and which has close and often illuminating relations with other fields such as statistics. Our approach is based on *regularization theory*; it is strictly related to the approximation technique called *Radial Basis Functions* and is equivalent to a certain class of multilayer networks.

One of the best ways to gauge how successfull our project has been is to look at the present activity that has originated from it. In our group and together with collaborators in Israel, Germany, England and Italy, we are now following four main directions of work.

2.2 Directions of Present Research

1. Developing the theory and related mathematical issues
2. Developing efficient algorithms for learning, including hardware implementations
3. Applying the technique to several problems such as:
   - Visual object recognition
   - Time-series analysis
   - Computer graphics
   - Autonomous navigation and control
   - Synthesis of early vision algorithms
4. Exploring possible implications for how the brain might work, and in particular:
   - How the brain may recognize 3-D objects
   - Whether simple, high performance visual tasks – such as hyperacuity tasks – depend significantly on a fast learning process.

3 Technical Milestones

We will first review the basic technique that we have developed and then describe some of our most recent results. Additional details can be found in the papers in the bibliography at the end, which list work done within our project on learning.

3.1 The HyperBF Technique

HyperBF networks (Poggio and Girosi, 1990, 1990a, 1990b, 1990c) are a class of feedforward networks with one layer of hidden units that compute functions of the form:

$$f(\mathbf{x}) = \sum_{\alpha=1}^{n} c_\alpha G(||\mathbf{x} - \mathbf{t}_\alpha)||_W) + p(\mathbf{x}) \tag{1}$$

where G is any conditionally definite positive function, $p(\mathbf{x})$ is a polynomial of low degree, W is a square matrix and $|| \cdot ||_{\mathbf{W}}$ indicates the following *weighted norm*:

$$||\mathbf{x}||^2_W \equiv W\mathbf{x} \cdot W\mathbf{x} \,. \tag{2}$$

The coefficients c_α, the "centers" $\mathbf{t}_\alpha$ and the matrix $\mathbf{W}$ are found during the learning stage, by minimizing a measure of the error between the network's prediction and each of the examples. After learning, the centers of the basis functions are similar to prototypes, since they are points in the multidimensional input space. Updating the centers during learning is therefore equivalent to modifying the corresponding prototypes, and corresponds to task-dependent clustering. Finding the optimal weights $\mathbf{W}$ for the norm is equivalent to transforming appropriately (for example, scaling) the input coordinates, and corresponds to task-dependent dimensionality reduction.

3.2 Theory

Our main line of investigation has been devoted to the problem of selecting an approximation technique, i.e., a specific network, because this is one of the choices that strongly influences the final performance. However, once an architecture has been chosen, there are other relevant problems that must be solved. One of these is related to the fact that in many cases the available data may contain outliers, and standard procedures (such as least square estimation) must be modified in this case. Here we show some results on these two topics.

3.2.1 Network Selection

Whenever we want to use some kind of network to solve a problem, two fundamental questions arise: a) how many hidden units are there? b) which activation function should the hidden units compute? We considered the first question under the assumption of an infinite number of examples. The number of units needed to approximate a function within a certain accuracy depends on the choice of the activation function, and on some characteristics of the function to be approximated, such as its dimensionality and degree of smoothness. For many classical spaces of functions and choices of the activation function, the dependence of the number of hidden units on the dimension is exponential, leading to the well-known phenomenon of "the curse of dimensionality". However, if some constraints are imposed on the target functions, better rates of convergence can be obtained. Using a result by Jones (1990) about the rate of convergence of iterative sequences in Hilbert spaces, we proved (Girosi and Anzellotti, 1991) that there exist classes of functions that can be approximated by a network of n radial units with an L_2 error of order $O(\frac{1}{\sqrt{n}})$. The dimension of the space influences the result only through a multiplicative constant, and the result is constructive in the sense that it shows an iterative algorithm that can achieve this rate of convergence. Similar results have been obtained by Barron (1991) for multilayer perceptrons, and this raises the question of the choice of the activation function, on which we did some experimental and theoretical work.

3.2.2 Network Selection: A comparison of MLP, HBF and Other Networks

Minoru Maruyama, Federico Girosi and Tomaso Poggio (1991a) have compared in numerical experiments several different activation functions, and therefore different techniques for learning from examples, considered as schemes for approximating multivariate functions from sparse data. In particular they considered multilayer perceptrons with one layer of sigmoidal hidden units, flexible Fourier series, multilayer perceptrons with exponential activation functions, Radial Basis Functions, and different forms of HyperBF networks. They have characterized their approximation performance (equivalent to generalization power) according to L_2 and L_∞ measures on sparse data from several different continuous functions of two and more variables, using several different training techniques. All the techniques, except that using exponential activation functions, performed well on average, and this led us to investigate possible relations between multilayer perceptrons and Generalized Radial Basis Functions (GRBF).

3.2.3 Network Selection: A Connection between MLP and HyperBF Networks

The main point of another project of Maruyama, Girosi and Poggio (1991b) has been to show that for normalized inputs, multilayer perceptron networks *are* radial function networks (albeit with a non-standard radial function). This provides an interpretation of the weights w as centers t of the radial function network, and therefore as equivalent to *templates*. This insight may be useful for practical applications, including better initialization procedures for MLP. Maruyama et al. also analyzed the relation between the radial functions that corresponds to the sigmoid for normalized inputs and well-behaved radial basis functions such as the Gaussian. In particular, they observed that the radial function associated with the sigmoid is an activation function that is good approximation to Gaussian basis functions for a range of values of the bias parameter. The implication is that a MLP network can always simulate a Gaussian GRBF network (with fewer parameters), but the converse is true only for certain values of the bias parameter. Numerical experiments indicate that the constraint is not always satisfied in practice by MLP networks trained with backpropagation. *Multiscale* GRBF networks, on the other hand, can approximate MLP networks with a similar number of parameters.

3.2.4 Dealing with Outliers

Given n noisy observations g_i of the same quantity f, it is common usage to give an estimate of f by minimizing the function $\sum_{i=1}^{n}(g_i - f)^2$. From a statistical point of view, this corresponds to computing the Maximum Likelihood estimate, under the assumption of Gaussian noise. However, it is well known that this choice leads to results that are very sensitive to the presence of outliers in the data. For this reason it has been proposed to minimize functions of the form $\sum_{i=1}^{n} V(g_i - f)$, where V is a function that increases less rapidly than the square. Several choices for V have been proposed and successfully used to obtain "robust" estimates. However, a justification and interpretation for their use is still lacking. We have shown (Girosi, 1991; Girosi, Caprile and Poggio, 1991) that for a class of functions V, which we call "effective potentials," using these robust estimators corresponds to assuming that our measures are affected by a

Gaussian noise whose variance is a random variable with a given probability distribution. Depending on the probability distribution of the variance of the noise, different shapes for V are obtained. Girosi (1991) gives characterization of the class of effective potentials in terms of positive definite functions in Hilbert spaces.

3.3 Parallel Algorithms

Learning the coefficients c_α, the $\mathbf{W}$ matrix and the $\mathbf{t}_\alpha$, that minimize an error functional of the L^2 type on the set of examples is a non-convex minimization problem. Gradient-descent is probably the simplest approach for attempting to find the solution to this problem. We have explored an even simpler optimization technique that can be successfully used to solve this class of problems (Caprile, Girosi and Poggio, 1991). Our algorithm combines aspects typical of many *genetic algorithms* with others typical of random descent techniques (Caprile and Girosi, 1990) into the concept of *adaptive noise.* We have tested the algorithm numerically in a variety of cases, and the results have been compared to the ones obtained by using a standard gradient descent with adaptive step technique. In all the cases considered, the best local minima were found by the nondeterministic algorithm, and preliminary experiments suggest that this may also hold true for a class of minimization problems wider than the one we have considered.

3.4 A Theory of How the Brain Might Work

We have proposed a quite speculative new version of the grandmother cell theory to explain how the brain, or parts of it, might work. In particular, we have analyzed how the visual system may learn to recognize 3-D objects. The model would apply directly to the cortical cells involved in visual face recognition. We have also outlined the relation of our theory to existing models of the cerebellum and of motor control. Specific biophysical mechanisms can be readily suggested as part of a basic type of neural circuitry that can learn to approximate multidimensional input-output mappings from sets of examples and that is expected to be replicated in different regions of the brain and across modalities. The main points of the theory are:

- The brain uses modules for multivariate function approximation as basic components of several of its information processing subsystems
- These modules are realized as HyperBF networks (Poggio and Girosi, 1990a,b)
- HyperBF networks can be implemented in terms of biologically plausible mechanisms and circuitry.

3.4.1 Fast Perceptual Learning in Visual Hyperacuity

We are beginning to apply the HyperBF technique to explain the fast acquisition of visual abilities in simple tasks from a few examples of the task. Tomaso Poggio, Manfred Fahle and Shimon Edelman (1991) were able to show that networks which solve specific visual tasks, such as the evaluation of spatial relations with hyperacuity precision, can be easily synthesized from a small set of examples using the HyperBF technique. They observe that in many different spatial discrimination tasks, such as determining the sign of the offset in a Vernier stimulus, the human visual system exhibits hyperacuity-level performance by evaluating spatial relations with the precision of a fraction of a photoreceptor's diameter. They propose that this impressive performance depends in part on a fast learning process that uses relatively few examples and occurs at an early processing stage in the visual pathway. They were able to show that this hypothesis is plausible by demonstrating that it is possible to synthesize, from a small number of examples of a given task, a simple (HyperBF) network that attains the required performance level. Then they verified with psychophysical experiments some of the key predictions of this conjecture. In particular, they proved experimentally that, quite surprisingly, fast stimulus-specific learning takes place in the human visual system and this learning does not transfer between two slightly different hyperacuity tasks. This may have significant implications for the interpretations of many psychophysical results in terms of neuronal models.

3.5 Applications

We have applied the HyperBF technique to several different domains:

- 3-D object recognition
- Synthesis of algorithms for early visual tasks, such as hyperacuity tasks
- Computer graphics
- Time-series analysis
- Adaptive control
- Indoor vision-driven autonomous navigation.

We briefly discuss two of of these applications.

3.5.1 Object Recognition

Edelman and Poggio (1990) applied the HyperBF technique to the problem of 3-D object recognition with promising results. They were able to synthesize a module that can recognize an object from any viewpoint after it learns its 3-D structure from a small set of 2-D perspective views, using the HyperBF network scheme. Their results were obtained with simulated wireframe objects, and assumed that the problems of feature extraction and matching were already solved. The problems of occlusions and spurious features were ignored. We have now successfully extended the technique to work with gray level images of real paper clips (Brunelli and Poggio, 1991).

It is interesting to mention that psychophysical experiments carried out on wire-frame objects and other objects confirm that "immediate" 3-D object recognition in humans seems to be based on a process of interpolation of 2-D views rather than the use of 3-D models.

We have also began to apply HyperBF networks to the problem of recognizing faces, using a small set of images of any given face as examples. This assumes that a few views for each person are available to train the network (our estimate for a generic 3-D object is between 20 and 100 2-D views). The theoretical low limit is two views (for the visible aspect) (Basri and Ullman, 1990;

Poggio, 1990a). We have therefore begun work aimed at characterizing how recognition from just *one* 2-D view may be accomplished if views of other ("prototypical") objects of the same class are available (Poggio, 1991). Clearly one single view of a 3-D object (if shading is neglected) does not contain sufficient 3-D information. If, however, the object belongs to a class of similar objects (prototypes) of which many views are known, it seems possible to make reasonable extrapolations and to guess correctly other views of the specific object from just one 2-D view of it. We are certainly able to recognize faces turned 20-30 degrees from the front from just one frontal view, presumably because we exploit our extensive knowledge of the typical 3-D structure of faces. At this point one can pose the following problem: *from one 2-D view of a 3-D object, generate other views, exploiting knowledge of views of other objects of the same class.* If this can be done, we can then use Poggio and Edelman's technique – and its extensions – by using the views we have generated as a training set. The point is to generate artificial examples of deformations for the specific object of interest by extracting information about allowed deformations from a set of examples of objects of the same class, using standard approximation techniques. Poggio (1991) discusses under which conditions and definitions of class this goal can be achieved.

3.6 Time-Series Analysis

Jim Hutchinson and Tomaso Poggio are engaged in the study of learning architectures, their parallel implementations, and their applications to large, real world problems in time-series prediction. The goals of this work are to investigate the potential of parallel implementations to help with problems of parameter estimation, handling of large problems, and use of previously intractable methods; to assess the applicability and usefulness of various learning networks to the problem of time-series prediction; to determine appropriate ways of achieving domain specific goals in time-series modeling, especially obtaining estimates of model fit (i.e., variance of outputs) and methods for iterating predictions; and to determine appropriate ways of handling domain specific problems in time series modeling such as limited sample size, embedding *a priori* structure into the learning architecture, and selecting and transforming useful inputs from a collection.

Results to date from this work are fairly preliminary. We have shown that the Radial Basis Function class of learning methods can be efficiently implemented on the Connection Machine to solving large problems. We have investigated various mechanisms for embedding the time-series prediction problem in the Radial Basis Function framework, and have preliminary results indicating that such systems outperform corresponding traditional linear models on an interesting class of financial time-series.

3.6.1 Early Visual Tasks

This technology has potential practical implications in terms of vision architectures that can learn from a set of examples to perform specific visual tasks such as inspection tasks, without explicit *ad hoc* programming.

4 List of Papers Relevant to the Project

References

[1] A.R. Barron. Universal approximation bounds for superpositions of a sigmoidal function. Technical Report 58, Department of Statistics, University of Illinois at Urbana-Champaign, Champaign, IL, March 1991.

[2] R. Basri and S. Ullman. Recognition by linear combinations of models. A.I. Memo No. 1152, Artificial Intelligence Laboratory, Massachusetts Institute of Technology, 1990.

[3] R. Brunelli and T. Poggio. Face recognition: features vs. templates. Technical Report 9110-04, I.R.S.T., Povo (IT), 1991.

[4] R. Brunelli and T. Poggio. Hyberbf networks for real object recognition. In *Proceedings of the International Joint Conference on Artificial Intelligence*, Sydney, Australia, 1991.

[5] R. Brunelli and T. Poggio. HyperBF networks for gender recognition. In *Proceedings Image Understanding Workshop*. Morgan Kaufmann, San Mateo, CA, January 1991b.

[6] B. Caprile and F. Girosi. A nondeterministic minimization algorithm. A.I. Memo 1254, Artificial Intelligence Laboratory, Massachusetts Institute of Technology, Cambridge, MA, September 1990.

[7] B. Caprile, F. Girosi, and T. Poggio. A nondeterministic method for multivariate functions minimization. In *Architetture Parallele e Reti Neurali*, Vietri sul Mare, Salerno, (IT), May 8-10 1991. (in press).

[8] S. Edelman and T. Poggio. Bringing the grandmother back into the picture: a memory-based view of object recognition. A.I. Memo 1181, Artificial Intelligence Laboratory, Massachusetts Institute of Technology, 1990.

[9] S. Edelman and T. Poggio. Models of object recognition. *Current Opinion in Neurobiology*, 1:270–273, 1991.

[10] F. Girosi. Models of noise and robust estimates. A.I. Memo 1287, Artificial Intelligence Laboratory, Massachusetts Institute of Technology, 1991.

[11] F. Girosi and G. Anzellotti. Rates of convergence of approximation by translates. A.I. Memo 1288, Artificial Intelligence Laboratory, Massachusetts Institute of Technology, 1991. (in press).

[12] F. Girosi and T. Poggio. Representation properties of networks: Kolmogorov's theorem is irrelevant. *Neural Computation*, 1(4):465–469, 1989.

[13] F. Girosi and T. Poggio. Networks and the best approximation property. *Biological Cybernetics*, 63:169–176, 1990.

[14] F. Girosi and T. Poggio. Regularization, radial basis functions, and recent extensions of networks for learning. In *Proc. of the 13th IMACS World Congress on Computation and Applied Mathematics*, Trinity College, Dublin, Ireland, July 1991.

[15] F. Girosi, T. Poggio, and B. Caprile. Extensions of a theory of networks for approximation and learning: outliers and negative examples. In R. Lippmann, J. Moody, and D. Touretzky, editors, *Advances in Neural information processings systems 3*, San Mateo, CA, 1991. Morgan Kaufmann Publishers.

[16] L.K. Jones. A simple lemma on greedy approximation in Hilbert space and convergence rates for projection pursuit regression and neural network training. *Annals of Statistic*, 1990. (to appear).

[17] M. Maruyama, F. Girosi, and T. Poggio. Techniques for learning from examples: Numerical comparisons and approximation power. A.I. Memo 1290, Artificial Intelligence Laboratory, Massachusetts Institute of Technology, 1991a. (in press).

[18] M. Maruyama, F. Girosi, and T. Poggio. A connection between HBF and MLP. A.I. Memo No. 1291, Artificial Intelligence Laboratory, Massachusetts Institute of Technology, 1991b. (in press).

[19] T. Poggio. A theory of how the brain might work. In *Proc. Cold Spring Harbor meeting on Quantitative Biology and the Brain*, 1990.

[20] T. Poggio. 3D object recognition: on a result by Basri and Ullman. Technical report # 9005–03, IRST, Povo, Italy, 1990a.

[21] T. Poggio. 3D object recognition and prototypes: one 2D view may be sufficient. Technical Report 9107–02, I.R.S.T., Povo, Italy, July 1991.

[22] T. Poggio. Economic models and time series: Learning from examples. *The Tactician, Citicorp Technology Office, New York*, 1991.

[23] T. Poggio and S. Edelman. A network that learns to recognize 3D objects. *Nature*, 343:263–266, 1990.

[24] T. Poggio, M. Fahle, and S. Edelman. Synthesis of visual modules from examples: learning hyperacuity. A.I. Memo No. 1271, Artificial Intelligence Laboratory, Massachusetts Institute of Technology, January 1991.

[25] T. Poggio and F. Girosi. A theory of networks for approximation and learning. A.I. Memo No. 1140, Artificial Intelligence Laboratory, Massachusetts Institute of Technology, 1989.

[26] T. Poggio and F. Girosi. Continuous stochastic cellular automata that have a stationary distribution and no detailed balance. A.I. Memo 1168, Artificial Intelligence Laboratory, Massachusetts Institute of Technology, 1990.

[27] T. Poggio and F. Girosi. Hyperbf: A powerful approximation technique for learning. In Patrick H. Winston and Sarah A. Shellard, editors, *Artificial Intelligence at MIT: Expanding Frontiers, Vol. 1*. M.I.T. Press, Cambridge, MA, 1990.

[28] T. Poggio and F. Girosi. A theory of networks for learning. *Science*, 247:978–982, 1990a.

[29] T. Poggio and F. Girosi. Networks for approximation and learning. *Proceedings of the IEEE*, 78(9), September 1990b.

[30] T. Poggio and F. Girosi. Extension of a theory of networks for approximation and learning: dimensionality reduction and clustering. A.I. Memo 1167, Artificial Intelligence Laboratory, Massachusetts Institute of Technology, 1990c.

[31] T. Poggio and F. Girosi. Learning algorithms and network architectures. In *Proc. Exploring Brain Functions: Models in Neuroscience*, Berlin, Germany, 1991. Dahlem Konferenzen.

[32] T. Poggio et al. The MIT Vision Machine. In Patrick H. Winston and Sarah A. Shellard, editors, *Artificial Intelligence at MIT: Expanding Frontiers, Vol. 2.* M.I.T. Press, Cambridge, MA, 1990.

[33] C. Weems, C. Brown, J. Webb, T. Poggio, and J. Kender. Parallel processing in the darpa strategic computing vision program. *IEEE Expert*, 6:23–38, 1991.